SciVerse ScienceDirect
Physics Procedia 24 (2012) 1406 – 1413
Physics Procedia

2012 International Conference on Applied Physics and Industrial Engineering

Application of Improved SAX Algorithm to QAR Flight Data

Yang Hui, Meng Fanxing

Dept. of Computer Science & Technology
Civil Aviation University of China
Tianjin 300300 CN

Abstract

During describing, storing and retrieving such operations on QAR flight data, traditional SAX can't overcome time series amplitude flex and timeline drift, so improved algorithm is proposed. QAR flight data will be divided into three stages and use algorithms to fill the cruise stage, thus allowing effective search for time series of different length. The experiment and item prove the feasibility and effectiveness. It increased greatly the efficiency of aircraft troubleshooting.

Keywords: SAX, QAR flight data, Similarity search, Fault diagnosis.

---

1. Introduction

Symbolic representation of time series presented in recent years, which is a discrete approach for time series data. It is a new method of information analysis by the theory of symbolic dynamics, chaos time series analysis and information theory, which can be to suppress noise excellently ，reduce the computing time and implement visualization easily. Symbol timing of the study experienced a gradual deepening of the process, But the most symbolic representation of all of the following fatal flaws: Time-series data is a typical high-dimensional data, most of the symbol timing can't achieve data compression; Symbol timing can't be defined in a distance-based and not be able to compare similar also, so the lack of practical value.

In the current symbolic timing of various methods, including the variance-based[1], entropy[2], hierarchical clustering[3], wavelet[4], symbol false neighbors method[5] and SAX(Symbolic aggregate approximation)[6].Dr. Eammon proposed the new method of SAX in 2003, Which is effectively to overcome the fatal flaw mentioned above for the first time and the performance is better or at least not lower than in other's way in many applications, therefore, it is subject to wide attention. Its main advantage is more simple and efficient than any other sign algorithm; In the process of symbolic realization, it achieved the data compression, noise reduction and ensure the lower bound Requirements that calculated the distance between two symbolic sequences In the symbol space, compared to the actual distance of the two time series, that will not be omitted. SAX approach has been widely applied in image processing, bioinformatics, data mining and machine learning. But it is only appropriate to follow the Gaussian distribution and limited variance in high density within the range of time-series data, because

doi:10.1016/j.phpro.2012.02.209

the approximate methods based on probability intervals is easy to lose some extreme information, such as financial time series data analysis[7].

Therefore, this paper proposed an improved SAX symbols algorithm to realize on the flight data preprocessing, in order to prepare similarity search[8] and apply to fault diagnosis[9].

2. Background and Related Work

Aircraft to complete a mission to go through taxi, takeoff, climb, cruise, down, approach and landing phases. Throughout the whole flight, the most complex operating phases are the take-off and landing phases, according to statistics, 68% of all aviation accidents occur in these two stages, therefore, on the study of the flight data, focuses on taxiing and take-off, climb, fall, approach and landing at these stages. Figure.1 selected four properties in different absolute height of QAR flight data, the vertical axis for the foot, abscissa in seconds.

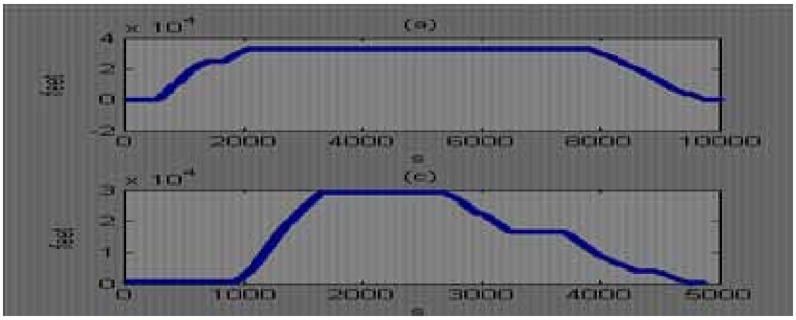


Figure .1 The absolute height of different flight

As can be seen from figure.1, despite the length of the flight data is different, the length of the taxi, takeoff, climb phase and decreased approach and landing phases are similar, Only large differences in the length of the cruise stage. As for the whole sequence searching in this type of flight data attribute, it will have the following questions:

1)Taking into account the face of large-scale time series database similarity search, the new representation can be efficient on the raw data compression, the new representation can efficiently compress the raw data and reduce data processing time and storage space.

2)How to improve the accuracy and fast of similarity search, thus improving the efficiency of fault diagnosis of aircraft.

3)Sequences set obtained by query algorithm, which contains all the similar sequences, without omission, that is, the complete query results.

3. Improved SAX algorithm

On these issues, there are two solutions: according to the stage, it divided the data, before the cruise stage is divided into the first stage, it is the third stage after the cruise stage; Another method is to convert the cruise phase which based correlation algorithm to the same character sequence length. For the previous method, due to the data of three stages are quite different in magnitude and the error is relatively large in the process of symbol, therefore, another way.

There are so many similar properties in the QAR flight data. selected height, the radio altitude, pressure altitude and ground speed correction, etc. For example, as shown in Figure.2:

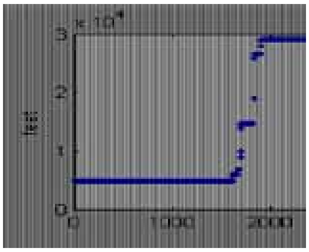


Figure.2(a) Selected altitude

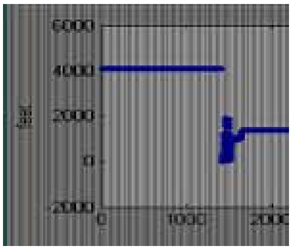


Figure.2(b) Radio height

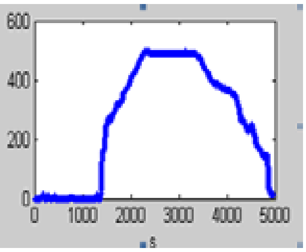


Figure.2(c) Barometric

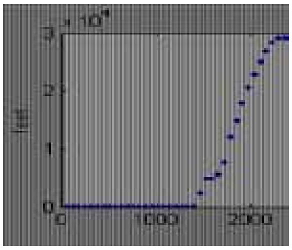


Figure.2(d) Ground speed correction altitude

When carrying out symbolic transformation in a similar property, such as radio height. First, it is to determine the number of data points by the method of PAA(Piecewise Aggregate Approximation)[10] in each paragraph, and then get the number of sub-paragraph, this will be a series of character sequence which are the data of similar to the first and third phases and different of the cruise stage; Second, according to the definition of overlap and similarity matching rate, it is to be found the location which will need to add characters to, add space to the location which equal the whole length sequence of characters; Finally, according to the related method, it is fill with the value of the location of space.

Figure.3(a), (b) shows an example, two sequences were expressed as T, S. First, the sequence T, S Will be the symbolic transformation, furthermore, the number of each piece contains 500 data points by PAA section, character set “a” value of 6. As shown in Figure.3,the sequence of characters which are defined as λ and μ were "aacdeeeeeeeeeeeedbaa" and "aabeeeeeeedcba", the value of “w” were 20 and 14.

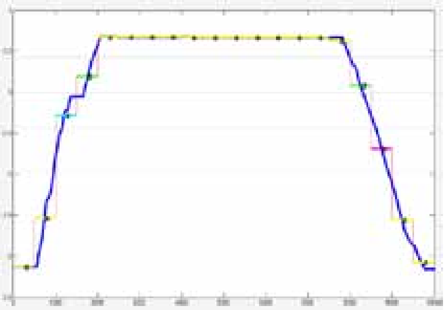


Figure.3(a) λ: before improvement
λ: aacdeeeeeeeeeeeedbaa

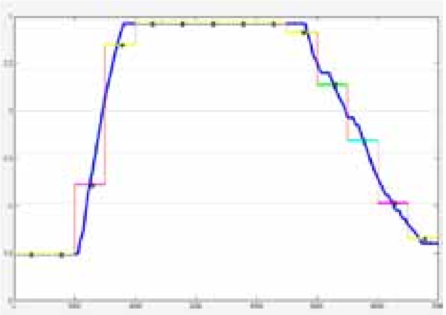


Figure.3(b) μ: before improvement
μ: aabeeeeeeeedcba

λ and μ compared in any position as shown in Figure.4:

Figure.4 λ and μ compared in any position

The length of a sequence of characters denoted by “n”(take the value of larger length of characters sequence, here, n=20), the number of same letters (a, e) denoted “m” (here, m = 8), the number of two overlapping strings denoted “r” (here r = 14).

Two equal length(add character to the shorter length of the sequences and equal to the longer) of characters sequence are matching, the number of overlapping characters divided by the length of character sequences is defined as the ratio of overlap:

$$OL=r/n \tag{1}$$

The number of character that is the same in corresponding positions divided by the length of character sequences, which is defined as matching rate:

$$MAT=m/n \tag{2}$$

The square of matching rate multiplied by the ratio of overlap, which is defined as similarity:

The square of matching rate multiplied by the ratio of overlap, which is defined as similarity:

$$Sim = MAT^2 * OL = (\frac{m^2}{n^2}) * \frac{r}{n} \tag{3}$$

In the case of the same as MAT, by the rate of overlap measured similarity, similarity is dependent on the matching rate, therefore, the impact of overlap for the similarity can be reduced. According to the definition of similarity, Just consider the maximum matching rate, that can get the maximum similarity.

The length of λ is defined as T_len, similarly, μ is defined as S_len (T_len>S_len), diff=T_len-S_len. When the last character of λ and first character of μ is in alignment, and get the similarity to Sim_1, μ shifted to the right one point, a calculation of the similarity index of Sim_2, when the first character of λ and the last character of μ is in alignment, and get the similarity to Sim_m(m=T_len+S_len-1), and then, in these similarity, calculated the maximum similarity: Sim_max=max{Sim_1,Sim_2,…,Sim_m}.

Corresponding to the two character sequence as shown in Table.1, the first row is λ and the first column is μ ,if the cross point of matrix elements match, marked '1 ', otherwise '0'. The slash through the cell that is the λ and μ match location and match situation, the first slash covered character sequences is S_len characters before λ and μ, in accordance with this continues, the total of slash is diff +1.

In order to add a space to the maximum matching rate, it must be to find a path with the most non-zero value in the T_len-S_len +1 article slashes. Because that move to a slash every time is equivalent to insert a space, the search path to follow the principles from left to right, when you insert a space, you can only match the back. Therefore, in accordance with from left to right, top-down principles, you can find a path with the highest non-zero value. in the diff +1 article slashes. As shown in Table.1:

TABLE.1 MATCHING TABLE

	a	**a**	**c**	**d**	**e**	**e**	**e**	**e**	**e**	**e**	**e**	**e**	**e**	**e**	**e**	**e**	**d**	**b**	**a**	**a**
a	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
a	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
b	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
e	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
e	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
e	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
e	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
e	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
e	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
e	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0
d	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
c	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
b	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
a	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1

The coverage by a slash do a conversion into the matrix R1, the first row of R1 represents the first slash of left in Table.1 , and so on. The matrix elements expressed the value through the cell values, in other words, the character is matched or not. In accordance with the above principles: from left to right, top to bottom, it will find a path with the most '1'. To facilitate that, the '1' replaced by the trekking number, said matrix R2.

$$R_1=\begin{bmatrix}1&1&0&0&1&1&1&1&1&1&0&0&0&0\\1&0&0&1&1&1&1&1&1&1&0&0&0&0\\0&0&0&1&1&1&1&1&1&1&0&0&0&0\\0&0&0&1&1&1&1&1&1&1&0&0&0&0\\0&0&0&1&1&1&1&1&1&1&0&0&0&0\\0&0&0&1&1&1&1&1&1&1&0&0&1&1\\0&0&0&1&1&1&1&1&1&1&1&0&0&1\end{bmatrix}\qquad R_2=\begin{bmatrix}1&1&0&0&1&1&1&1&1&1&0&0&0&0\\2&0&0&2&2&2&2&2&2&2&0&0&0&0\\0&0&0&3&3&3&3&3&3&3&0&0&0&0\\0&0&0&4&4&4&4&4&4&4&0&0&0&0\\0&0&0&5&5&5&5&5&5&5&0&0&0&0\\0&0&0&6&6&6&6&6&6&6&0&0&6&6\\0&0&0&7&7&7&7&7&7&7&7&0&0&7\end{bmatrix}$$

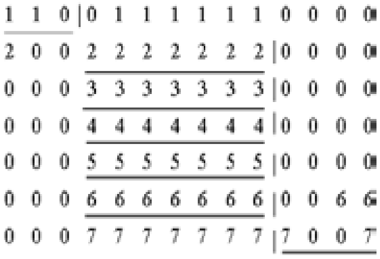
The number of non-zero value expressed as the amount of information, as follow from left to right, top-down principles, the row i can contain the amount of information in. the row i+1,in other words, if you selected part of the elements in these two lines, you can make maximum amount of information in row i. From the penultimate row to the first row in matrix R2 , the first line contains the maximum amount of information in all the lines, that is, to find the optimal path. Using brute-force method to find the optimal division(select the left element in the first line and the right element in row i+1 to the maximum amount of information). Algorithm steps are as follows:

Step 1.Take an element of the penultimate line, and then take S_len-1 elements in the last line, calculating the maximum amount of information is 8;

Step 2.Take two elements of the penultimate line, and then take S_len-2 elements in the last line, calculating the maximum amount of information is 8;

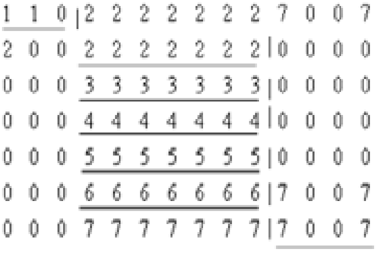
Step 3.Followed by exhaustive calculation, the division of the largest information is to take the first 10 elements in the penultimate line and take the last four elements in the last line, the amount of information is 9;

Step 4.From back to front, repeat steps 1,2,3 until the first row, the results of all of the division as shown in Figure.5:



1	1	0	0	1	1	1	1	1	1	0	0	0	0
2	0	0	2	2	2	2	2	2	2	0	0	0	0
0	0	0	3	3	3	3	3	3	3	0	0	0	0
0	0	0	4	4	4	4	4	4	4	0	0	0	0
0	0	0	5	5	5	5	5	5	5	0	0	0	0
0	0	0	6	6	6	6	6	6	6	0	0	6	6
0	0	0	7	7	7	7	7	7	7	7	0	0	7

Figure.5 the division



1	1	0	2	2	2	2	2	2	2	7	0	0	7
2	0	0	2	2	2	2	2	2	2	0	0	0	0
0	0	0	3	3	3	3	3	3	3	0	0	0	0
0	0	0	4	4	4	4	4	4	4	0	0	0	0
0	0	0	5	5	5	5	5	5	5	0	0	0	0
0	0	0	6	6	6	6	6	6	6	7	0	0	7
0	0	0	7	7	7	7	7	7	7	7	0	0	7

Figure.6 the division after finishing

Step 5. Set the elements zero in the upper right corner of division. elements in the lower right corner will replace the upper right corner, and retain the elements in the upper left corner, the division as shown in Figure.6.

Step 6. Add a space where the values are increasing, the number of spaces equal to the difference between the values and after the change (except zero elements). As shown in Figure.7.

λ= aacdeeeeeeeeeeeeedbaa

μ= aab_eeeeeeee____dcba

Figure.7 the maximum matching similarity

As such filling method, the cruise phase after the symbolic obtained is generally the same consecutive characters, according to the method of filling vacancies, such as the use of the nearest filling method. After filling, μ is “aabbeeeeeeeeeeeeedcba”, and then it indexed for search by the index method.

The improved SAX algorithm, first, time series data was carried out flexible treatment; and then calculating the maximum similarity ensure the accuracy and effectiveness in similarity search; finally, by exhaustive method it is to find a similar division to ensure the completeness of search results.

4. Experimental environment and data

Hardware Environment: Intel (R) Core (TM) 3.00GHZ, memory.: 1.96GB.

Software Environment: The operating system is Windows XP, the software of MATLAB7.0.

Data selection and preprocessing: This data is from the airborne QAR data, the model is the Boeing 737,the flight data is from August 2008 to August 2009.

Similarity Search: During the similarity search, because the fault model provides is limited the airlines, This part of the fault data is as the fault model data. As shown in Figure.8 ,the two groups of attribute columns, in accordance with its corresponding analog fault models, such as Figure.8(a), (b), (c) and (d) is similar to a group of fault data, but the fault location and length of the data points are different. When creating the fault model, the fault model included the fault name, symbol of the sequence, fault sources, fault description and fault type. Fault category are EICAS Message, Observed Fault, Cabin Fault and Maintenance Message; The similarity search to retrieve the fault information includes name and QAR file name; File information includes the QAR file name, flight number, tail number, flight date, departure time, landing time, origin and destination.

In the process of indexing, first, determine the longest length of the data in QAR flight data which need for similarity search, and get the transformation of the SAX symbol, the number of sub-section is the length of time series data divided by 500(here, w=28),the size of symbol set "a" is 6; Then by the SAX improved algorithm, symbolic sequence is filled and create the index file for similarity search, the results are shown in Table.2.

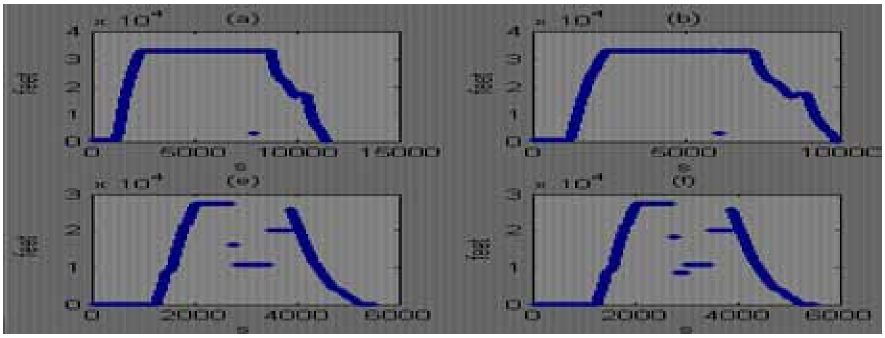


Figure.8 the fault data of absolute height

Table.2 by the index method, it obtained the experimental data which is approximate search and accurate search for fault model. the first row is the result of approximate search and the second row is the result of accurate search. Index file is initialized by the time of creating the index tree, the number of index file is the number of leaves' nodes in the index tree.

TABLE.2 APPROXIMATE / ACCURATE SEARCH

Search method	Index file initialization	Number of index files	Find time	Distance calculation	Accuracy
Approximate search	0.056ms	53	0.218ms	-	0.863
Accurate search	0.056ms	53	6.174ms	243.472	0.901

The experiment results show that in both cases, although the find time need a long time in accurate search, it greatly increased the accuracy rate, especially, it is very important for aircraft fault diagnosis. But it is only appropriate to follow the Gaussian distribution and limited variance in high density within the range of time-series data, when the length of time series data is large difference, this method has some errors.

The results show the similarity search algorithm can obtain the correct diagnosis results, and that the method can be done fault location by similarity search. However, because of the experimental conditions, the fault model is not enough and the lack of large time series databases, the algorithm of the time series similarity search in aircraft fault diagnosis Still needs to be verified.

5. Conclusions and Future Directions

The improved SAX algorithm used to data mining for QAR flight data and troubleshooting them In this paper, and analyzes its efficiency in the search and calculate the accuracy of search results, greatly improved the efficiency of troubleshooting. However, the data segmentation algorithm is more flexible, making search results more sensitive to noise data and the search results is not very satisfactory, the more sub-associate and the more loss of information, the accuracy of the results may also have been affected.

6. Acknowledgment

This work is supported by the project of National Natural Science Foundation of China (No.60776806) and the project of Science and Technology Foundation for the Civil Aviation of China (No.MHRD200806)

References

[1] Xu Z S, Wei C P. A consistency improving method in the analytic hierarchy process. European J of Operational Research , 1999 , 116 (2) : 443-449.

[2] Orlovsky S A. Decision-making with a fuzzy preference relation[J]. Fuzzy Sets and Systems , 1978 , 1 (3) : 155-167.

[3] Kacprzyk J . Group decision making with a fuzzy linguistic majority[J]. Fuzzy Sets and Systems , 1986 ,18 (2) : 105-118.

[4] Chiclana F , Herrera F , Herrera Viedma E , et al. A classification method of alternatives for multiple preference ordering criteria based on fuzzy majority[J].J of Fuzzy Mathematics , 1996 , 4 (4) : 128-143

[5] Van Laarhoven P J M , Pedrycz W. A fuzzy extension of satty'spriority theory [J]. Fuzzy Sets and Systems ,1983 , 11 (1) : 229-241.

[6] Jiang Y P , Fan Z P. A practical ranking method for reciprocal judgment matrix with triangular fuzzy numbers[J]. Systems Engineering , 2002 , 20 (2) : 89-92.

[7] Gong Z W , Liu S F. Consistency and priority of triangular fuzzy number complementary judgment matrix[J]. Control and Decision , 2006 , 21 (8) : 903-907.

[8] AGRAWAL R,FALOUSTOS C,SWAMI A.Efficient similarity search in sequence Databases[C]//Proceedings of 4th International Conference on Foundations of Data Organization and Algorithms Berlin: Springer,1993,69-84.

[9] Air Transport Systems Press Releases:www.sac.honeywell.com/atsrelhtml,1996

[10] Eamonn Keogh,Kaushik Chakrabarti,Michael Pazzani and Sharad Mehrotra Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases. Knowledge and Information Systems,(2001) 3:263-286